Local Model of the Air Quality on the Basis of Rough Sets Theory

Filip Mezera and Jiří Křupka

Institute of System Engineering and Informatics, Faculty of Economics and Administration,
University of Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic
st5360@student.upce.cz, jiri.krupka@upce.cz

Abstract. This article deals with the air quality modelling in two selected localities in the Czech Republic (CR). Data for the modelling were gained from the public sources. Primary source was the data server Czech Hydro Meteorological Institute (CHMI). Rough set theory (RST), Decision Trees (DTs) and Neutral Networks (NNs) were used for the analysis and the results comparison. At the end of the article there is the possible usage of the outputs of the models described. Outputs can help with the health protection of the inhabitants through the regulations set by the public administration authorities.

1 Introduction

Risks connected to the polluted air are the ones of the main environmental dangers [26,31], which are solved not only by regions, countries, but also by international organizations. This article deals with the synthesis and analysis of the air quality model in the selected localities of Czech Republic (CR). The model is aimed at dust particles (PM10) and weather character in two selected localities CR – Pardubice and Ostrava's neighbourhood. PM10 were selected, because they create an important part of the air quality and then they also carry the dangers of respiration diseases. Above all, small children can suffer from asthma or chronic inflammation of the upper respiratory tract [19]. Dust particles also carry carcinogenic substances which make higher the cancer risk significantly [2,3,28,31].

Using of methods of artificial inteligence in the weather conditions has been described in many articles, for example in [4]. Problematic of the air quality modelling in the CR regions was solved in the years 2007-2011 within the National Programme of Research of Ministry of Environment CR "The environment and natural resources protection". It was a project no. SP/4i2/60/07 titled "Indicators for Valuation and Modelling of Interactions among Environment, Economics and Social Relations". Suggested models used with data [21,22] from mobile and stationary meteorology stations in Pardubice's neighbourhood. Theories of Neutral Networks (NNs), fuzzy sets (FSs), Rough Sets Theory (RST) and Decision Trees (DTs) were used for creating the classificatory models [21,22].

Particular situation in every place is dependent on local conditions [6]. There are three main sources which participate at the overall level of the air pollution (pollution). First of them is the pollution from the big stationary sources (such as heating plants, power stations, ironworks etc.). Overall level of the pollution from these

V. Snasel et al. (Eds.): SOCO Models in Industrial & Environmental Appl., AISC 188, pp. 277–286.
springerlink.com

sources is considered to be quite stable. Second source is the local heating. This source is distinctly dependent on the type of fuel which is used in the selected locality and then also on the weather. In the time of the decreasing temperatures the consumption of fuel and also the amount of substances in the air increases. The last part is the vehicular traffic, which significantly fluctuate between the main rush hours [2]. The public administration authorities can react directly with the effective remedies to decrease the level of pollution. E.g. in the years 2006, 2008, 2009 and 2010 were announced "regulation states" for the Moravian-Silesian region connected to the smog in the locality. This problem was, however, struggled in the big part of CR on 15th Nov. 2011, including Pardubice's region. Smog situation is defined in the regulation §8 art. 1 of the Act [32] as „state of the exceptional air pollution, where the level of air pollution by the polluting substance exceed the particular limit set by the implemented regulation" explained, with reference to the article 1, that „the exceptional limit is thought to be the level of the air pollution, while exceeded, where is a risk of health harm or harm of the ecosystem in very short exposition time" (regulation §8 art. 2 of the Act). The limit is 50µgm^{-3}.

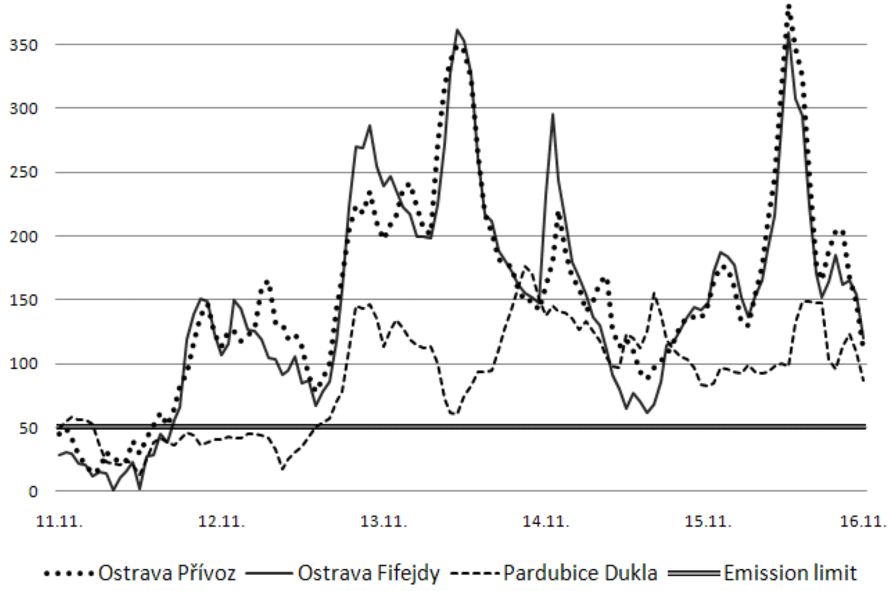

Fig. 1. Five-day concentration of PM10 in Ostrava (Přívoz and Fifejdy) and Pardubice (Dukla)

At Fig. 1 is possible to see that the limits in Pardubice are slightly exceeded (13th Nov. 2011). This did not lead to the signal of warning.

The warning is signalled unless the concentration of the suspended particles PM10 exceeds the limit 100 µgm^{-3} in average within the last 24 hours. In Ostrava (measuring stations Přívoz and Fifejdy) at that time reached values of 200 µgm^{-3} and more. Situation in Pardubice was quite worse during the afternoon on 12th Nov. 2011, when it momentarily exceeded the value of 140 µgm^{-3}. Values leading towards the warning signal were reached on 13th Nov. (3:00pm), when the average in the last 24 hours was

101.52 μgm^{-3}. Measured values are accessible with 3 hours delay on the CHMI portal [7]. But information about regulation state was publicly accessible with more than 18 hours delay. Pardubice's town hall published the smog announcement [18] the following day, i.e. 14th Nov. 2011. The next source of information for the inhabitants is regional, especially public broadcasting. That informed about the situation on 14th Nov. 2011 (10:36am) [5]. This situation does not correspond to the inhabitants needs. Mainly for seniors and young children exposed to the values higher than 100 μgm^{-3} (double as high as allowed norms) is harmfull. At the present delay of system of warning can reach up to two days, from the first contravention of level of 100 μgm^{-3}.

One of the options in an increase of the present state is the proceeding the air quality model, which can be used for prediction of the future state. It is possible to use the measured values of PM10, as well as the weather character. This has a huge impact to the air quality pollution [6]. If the model shows the high rate of accuracy, it could be possible to use it for informing the inhabitants.

2 Problem Formulation

Moravian-Silesian Region (Ostrava is its part) is the unique part in concentration of the large stationary sources of pollution (sources included to databases REZZO 1 and 2). This part of the pollution does not have any momentary differences. They form the local level of the background, thanks to which the states of the pollution are better identified. Due to the often occurrence of smog situations, Ostrava's neighbourhood is suitable for creation of the general model valid for the other CR regions. Air quality is monitored from the long-lasting point of view and in extreme situations can be regulated. Their portion of pollution is between 30 to 50 % in Ostrava [8]. Local sources (households, small polluters) have significant portion at the situation in the local neighbourhood and their regulation is very difficult. There is an assumption of correlation between the amount of giving off combustion products and the weather (especially the temperature). The portion at the pollution is 30-50 % in Ostrava, in the case of Pardubice it is estimated to 50 %. Other sources, such as traffic, can be regulated as well as momentarily and long-lastingly (travel by public transport for free, restrictions in entrance to problematic locations etc.). Their portion is between 15-40 % [2,3,8].

The weather plays the important role within the short-term air quality [8]. At the large sources we assume the independence in amount of harmful substances given off towards the weather [8]. At the small sources and traffic the correlation with the temperature characteristics can be estimated [2]. Moreover, the inverse character of the weather and wind speed will influence negatively the air quality. They cause that the harmful substances accumulate at the exposed places [10].

Considering the low number of measuring stations in the city with lower normal level of pollution [6], as they are Pardubice or Hradec Králové, the model must be counted with specific level of generalization. In every locality it is then possible to monitor certain differences of maximal pollution values in the time [13]. Considering levels of pollution, which are used for warning and regulation, the differences need not to be reflected in the model.

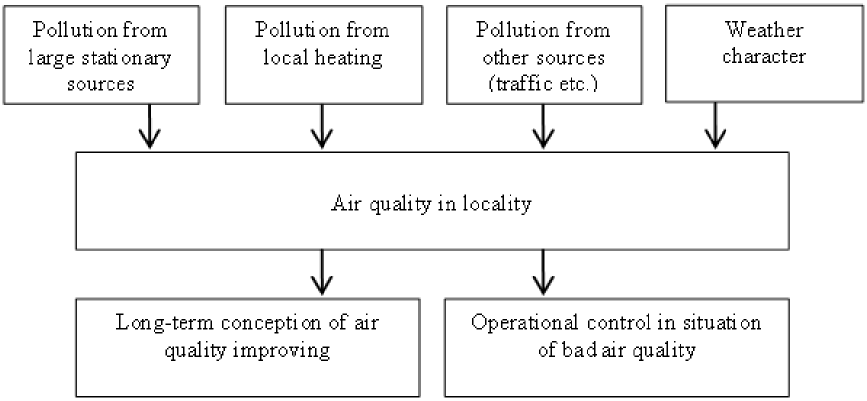
Pollution from large stationary sources
Pollution from local heating
Pollution from other sources (traffic etc.)
Weather character
Air quality in locality
Long-term conception of air quality improving
Operational control in situation of bad air quality

Fig. 2. Model of air quality in the selected location

2.1 Data Description and Pre-processing

The values measured at certain areas function as indicators of air quality. Data about the weather are gained from meteorological stations at the airports Ostrava – Mošnov and Pardubice. Moreover, the inputs from the stationary stations measuring the pollution were used. For Ostrava there were stations: Bartovice, Českobratrská, Fifejdy, Mariánské Hory, Poruba, Přívoz, and Zábřeh. In Pardubice there is one station measuring PM10 and that is at Dukla. Every factors and their impact to specific air pollution situation can be examined by modelling. In Pardubice the examination was done and outputs are in [12]. The examination was pursued with accurate data corresponding to the local conditions. Situation with excessive amount of PM10 is not typical in Pardubice. It is assume, that the model specified for Moravian-Silesian Region would be calibrated gradually to the local conditions. The model uses 17 input variables, which are described in the following Table 1.

For the dust particles classification the attributes l_1, l_2, m_1, m_2, …, m_{11}, m_{12} are used. In the years 2006 to 2011 there was overall 451 observations selected. Training and testing sets contained 366 observations altogether from the year 2008. Validation set was created by the data from the years 2006, 2009, 2010 and 2011, when the air pollution limits were contravened. Data from the years 2007 and 2008 were not accessible.

When model was created, the results mentioned in [12] were evaluated. The variables, which are relevant for the PM10, were selected for next treatment. The variables l_1, l_2, c_1, c_2, m_9 and m_{10} were discarded, because they seemed to be a little relevant on the level of used generalization. From the parameters m_1 to m_6 were then derived 3 new variables d_1, d_2 and d_3, which record the occurrence of the inverse character of the weather. Then the scale of air quality mentioned by CHMI was also used [6], Table 2. The derived variable d_4 that describes the PM10 status in the time period t (i.e. minus 24 hours) was assigned by categories.

Table 1. Selected input variables of the model of the dust particles amount PM10

Name of variable	Variable	Type of variable
Station	l_1	set
Type of station	l_2	set
Day in the year	c_1	discrete
Day of the week	c_2	discrete
Average/Maximal/Minimal day temperature city	$m_1/m_2/m_3$	continuous
Average/Maximal/Minimal day temperature Lysá Hora	$m_4/m_5/m_6$	continuous
Average/ Maximal wind speed	m_7/m_8	continuous
Wind direction in the morning/in the afternoon	m_9/m_{10}	discrete
Air humidity/pressure	m_{11}/m_{12}	continuous
Difference between the average/maximal/minimal day temperature in the city and at Lysá Hora	$d_1/d_2/d_3$	continuous
PM10 in the time (last 24 hours average)	d_4	continuous

Table 2. Scale of air quality

Index	Air quality	PM10 (average value per 1h in μgm^{-3})
1	Very good	0 – 15
2	Good	>15 – 30
3	Satisfactory	>30 – 50
4	Convenient	>50 – 70
5	Bad	>70 – 150
6	Very bad	>150

3 Suggestion and Model Analysis

Design of the model at the basis of RST is shown in Fig. 3. RST [1,15,24,25] is based on searching common features from the data. It works with uncertainty within the upper and lower approximation and boundary region. Design of the model follows the work [12,13], which confirmed the ability of RST usage during the air quality modelling.

Categorized variables $\{k_1, k_2, k_3, k_4, k_5, k_6\}$ (Table 3) were derived from the former variables $\{m_7, m_8, d_1, m_{11}, m_{12}, d_4\}$ (Table 1) for the needs of RST usage thanks to the equidistant scaling, except k_6. Their values and amounts of categories were set experimentally or come out of the description of the phenomenon due to [6]. Estimated parameter v_y an average value of PM10 in the following 24 hours had only two categories. Results of RST model were compared with DTs (algorithms C5 a CRT) [14] and NNs. For algorithms DTs and NNs were used former continuous variables described in Table 1.

There were 23 rules generated for determining the output parameter of the model v_y by software Rough Sets Exploration System (RSES) [29]. The rules were generated using the algorithm Learning from Examples Module 2 (LEM2) [9]. Obtained rules show the importance of each attribute. The most important parameters are inversion k_3, wind speed k_1 and average amount of PM10 in the last 24 hours k_6. All ten rules

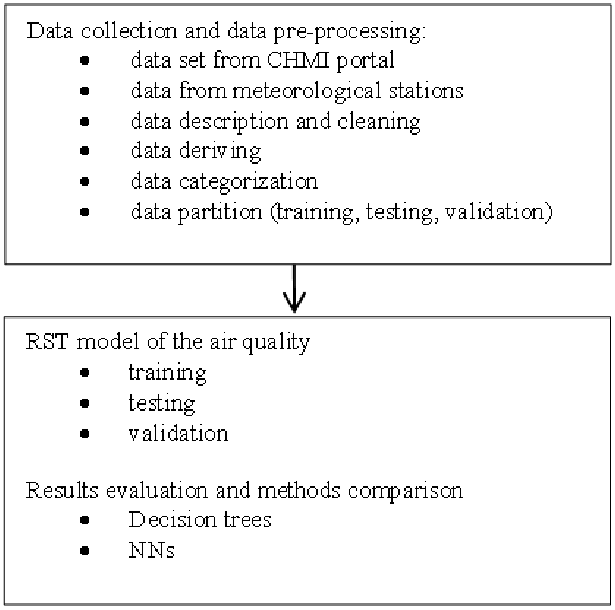
Data collection and data pre-processing:
data set from CHMI portal
data from meteorological stations
data description and cleaning
data deriving
data categorization
data partition (training, testing, validation)
RST model of the air quality
training
testing
validation
Results evaluation and methods comparison
Decision trees
NNs

Fig. 3. Suggestion of the model via RST and its analysis

consist of these parameters which determine the negative value of v_y , i.e. $v_y=1$. The most common negative rule is:

$$\text{IF } k_1=0 \text{ AND } k_3=1 \text{ AND } k_6=2 \text{ THEN } v_y=1. \tag{1}$$

It means that, in case of the inverse character of the weather and slow wind speed, the air pollution stays concentrated at one place and makes the air quality worse in the selected locality.

Table 3. Categorized variables for RST

Attribute / Name of atribute	Values of attributes (0; 1; 2; 3; 4)
k_1 / Day average wind speed [ms^{-1}]	0 is <9; 1 is 9–13; 2 is 13–17; 3 is 17–21 and 4 is >21
k_2 / Maximal wind speed [ms^{-1}]	0 is < 9; 1 is 9–14; 2 is 14–19; 3 is 19–24; 4 is >24
k_3 / Inverse weather character [ºC/100 above the sea level]	0 is <2; 1 is 2–5; 2 is >5
k_4 / Humidity [%]	0 is <66; 1 is 66–76; 2 is 76– 86; 3 is >86
k_5 / Pressure [hpsc m^{-1}]	0 is <1005; 1 is 1005–1012,5; 2 is 1012.5–1020; 3 is >1020
k_6 / Average amount of PM10 in the last 24 hours [µgm^{-3}]	0 is <22; 1 is 22–37; 2 is 37– 70; 3 is >70
v_y / Average amount of PM10 in the following 24 hours [µgm^{-3}]	0 is <70; 1 is >70

Accuracy of prediction reached 96.4 % (Table 4).

Table 4. Confusion matrix

		Predicted value	
		0	1
Detected value	0	65	2
	1	1	15

The result 96.4 % was compared with the results of DTs [13] and NNs. From the suggested DTs there were two, which have the accuracy higher than 90 %. One of them used the algorithm C5-boost and the second the algorithm CRT-boost. NNs with the good results were Multi Layer NN (MLP) and Radial Basis Function (RBF) NN. MPL NN has 6 neurons in the hidden layer and RBF NN 10 neurons. Comparison of the model results is in Table 5.

The verification of the results was done at the validation set (Table 5). Its specificity was much higher ratio of the days with negative air quality. This results in lower accuracy of the examined methods. Prediction ability of the RST and algorithm C5-boost and NNs methods is still quite high. Considering the complexity of the calculation and the following interpretation of the results, RST seems to be more robust and suitable for making a prediction model of the PM10 in the air.

There were not found the other considerable differences at the weather impact to the smog situation in Ostrava (accuracy 90.8 %) and Pardubice (accuracy 89.9 %). There is still the most important parameter the value of PM10, the inverse character of the weather and the average wind speed.

Table 5. Comparison of the results of the models at the basis of RST, DTs and NNs

Method	Accuracy (test set) [%]	Accuracy (validation set) [%]
RST	96.4	90.7
C5-boost	96.5	89.4
CRT-boost	93.8	84.7
MLP	96.2	91.3
RBF	95.2	90.8

If there were used the calculated rules at the basis of RST towards the predicted value v_y “Average amount of PM10 in the following 24 hours” during the warning against the harmful impact of the smog situation, it could be possible to shorten the length of the delay from the present average 36 hours to 3 hours. In that time are verifing data onto the CHMI server.

4 Conclusion

In this article there was introduced the problematic of the air quality and the solution of the warning system of the inhabitants against the negative impacts, which are connected to the stay at the exposed places. Nowadays, there is a average of 36 hours delay in the informing the public about the air pollution. This delay is too long, mainly with respect to the special groups (children, older people and people with respiration diseases). Creation the model, which identifies the bad quality of the air with the outlook of the following 24 hours, is the possibility how to prevent the selected groups of inhabitants from the health problems. Models are oriented to the amount of dust particles PM10 and the occurrence of smog in Ostrava and Pardubice's neighbourhood.

Firstly, it was necessary to find out, process and describe the data from the meteorological and air pollution stations. At the same time we have characterized both examined regions. RST work with categorized variables, therefore there were derived six variables. Data were then divided into three categories - training, testing and validating. Division of training and testing data was made by using the software IBM SPSS Modeler 14.2. For the rules calculation the software RSES 2.2.2 was used. The result rules were used for the prediction of "Average amount of PM10 in the following 24 hours".

Program RSES proved to be a suitable program for rules creation. The quality of the prediction at the tested set reached 96.4 %. Then there was a comparison of DTs and NNs made. The number of generated rules pointed out how robust RST is. During the validation by RST there were very good results reached (RST 90.7%, DTs 89.4% and MLP 91.3%).

Problems with air pollution belong to the important parts of the sustainable development of the regions from the long-lasting point of view. At the present trend of the larger cities development the solution and optimization of the air pollution will be the priority. The model of testing in other large cities of CR seems to be a good idea. With reference to the used methods the other high tech equipment is able to be used, e.g. rough-fuzzy access, case-based reasoning etc.

Acknowledgments. This work was supported by the project No. CZ.1.07/2.2.00/28.0327 Innovation and support of doctoral study program (INDOP), financed from EU and Czech Republic funds.

References

[1] Aviso, K.B., Tan, R.R., Culaba, A.B.: Application of Rough Sets for Environmental Decision Support in Industry. Clean Technologies and Environmental Policy 10, 53–66 (2007)

[2] Bellander, T., et al.: Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. Environmental Health Perspectives 109(6), 363–369 (2001)

[3] Brauer, M., et al.: Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children. American Journal of Respiratory and Critical Care Medicine 166, 1092–1098 (2002)

[4] Corchado, E., Arroyo, A., Tricio, V.: Soft computing models to identify typical meteorological days. Logic Journal of the IGPL 19(2), 373–383 (2011)

[5] ČRo Pardubice (Český rozhlas): Pardubicko trápí smog (2011), `http://www.rozhlas.cz/pardubice/zpravodajstvi/_zprava/975676` (accesed November 17, 2011)

[6] Český hydrometeorologický Ústav (2012), http://www.chmi.cz (accesed February 1, 2012)

[7] Český hydrometeorologický ústav: Data AIM v grafech (2012), http://pr-asv.chmi.cz/IskoAimDataView/faces/aimdatavw/viewChart.jsf (accesed February 1, 2012)

[8] Černikovský, L., Volný, R.: Znečištění ovzduší a jeho zdroje v Ostravě. In: Konference o kvalitě ovzduší v Ostravě (April 2, 2012), http://www.ostrava.cz/cs/o-meste/zivotni-prostredi/6.-konference-o-kvalite-ovzdusi-v-ostrave-2012 (accesed April 15, 2012)

[9] Grzymala-Busse, J.W., Wang, A.Z.: Modified algorithms LEM1 and LEM2 for Rule Induction from Data with Missing Attribute Values. In: 5th Int. Workshop on Rough Sets and Soft Computing (RSSC 1997) at the Third Joint Conference on Information Sciences (JCIS 1997), pp. 69–72. Research Triangle Park, NC (1997)

[10] Horák, et al.: Bilance emisí znečišťujících látek z malých zdrojů znečišťování se zaměřením na spalování tuhých paliv. Chemické Listy 105, 851–855 (2011)

[11] IBM SPSS Modeler 14.2 User's Guide (2012), ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/Users-Guide.pdf (accesed January 17, 2012)

[12] Jirava, P., Křupka, J., Kašparová, M.: System Modelling based on Rough and Rough-Fuzzy Approach. WSEAS Transactions on Information Science and Applications 10(5), 1438–1447 (2008)

[13] Kasparova, M., Krupka, J., Jirava, P.: Approaches to Air Quality Assessment in Locality of the Pardubice Region. In: 5th Int. Conf. Environmental Accounting Sustainable Development Indicators (EMAN 2009), Prague, Czech Repulbic, pp. 1–12 (2009)

[14] Kasparova, M., Krupka, J.: Air Quality Modelling by Decision Trees in the Czech Republic Locality. In: 8th WSEAS Int. Conf. on Applied Informatics and Communications (AIC 2008), pp. 196–201. WSEAS Press, Greece (2008)

[15] Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In: Pal, S.K., Skowron, A. (eds.) Rough-Fuzzy Hybridization: A New Trend in Decision-Making, pp. 3–98. Springer, Singapore (1998)

[16] Kudo, Y., Murai, T.: A method of Generating Decision Rules in Object Oriented Rough Set Models. In: Rough Sets and Current Trends in Computing (RSCTC 2006), Kobe, Japan (2006)

[17] Maimon, O., Rokach, L.: Decomposition metodology for knowledge discovery and data mining. World Scientific Publishing, London (2005)

[18] Město Pardubice – Město je ohroženo smogem, byl vyhlášen signál upozornění (2011), http://www.pardubice.eu/urad/radnice/media/tiskove-zpravy/tz2011/tisk-111114.html (accesed November 19, 2011)

[19] Nařízení vlády 350/2002 Sb., kterým se stanovují imisní limity a podmínky a způsob sledování, posuzování, hodnocení a řízení kvality ovzduší, v platném znění (2002)

[20] Neri, M., et al.: Children's exposure to environmental pollutants and biomarkers of genetic damage: II. Results of a comprehensive literature search and meta-analysis. Mutation Research/Reviews in Mutation Research 612(1), 14–39 (2006)

[21] Olej, V., Obršálová, I., Křupka, J. (eds.): Modelling of selected areas of sustainable development by artificial intelligence and soft computing: regional level. Grada Publishing, The Czech Republic (2009)

[22] Olej, V., Obrsalova, I., Krupka, J. (eds.) Environmental Modeling for Sustainable Regional Development: System Approaches and Advanced Methods. IGI Global (2011)

[23] Pal, S.K., Skowron, A. (eds.): Rough-Fuzzy Hybridization: A New Trend in Decision Making. Springer, Singapore (1999)

[24] Pawlak, Z.: Rough Sets – Theoretical Aspects of Reasoning about Data. Kluwer, Boston (1991)

[25] Pawlak, Z.: Rough set approach to knowledge-based decision support. European Journal of Operational Research 99, 48–57 (1997)

[26] Portney, R.P., Stavins, R.N.: Public Policies for Enviromental Protection, Washington (2000)

[27] Rokach, L., Maimon, O.: Data mining with decision trees: Theory and applications. World Scientific Publishing, London (2008)

[28] Topinka, J., Binková, B., Mračková, G.: Influence of GSTM1 and NAT2 genotypes on placental DNA adducts in an environmentally exposed population. Environmental and Molecular Mutagenesis 30, 184–195 (1997), doi:10.1002/(SICI)1098-2280(1997)30:2<184::AID-EM11>3.0.CO;2-9

[29] Skowron, A., Bazan, J., Szczuka, M.S., Wroblewski, J.: Rough Set Exploration System (version 2.2.2) (2009), http://logic.mimuw.edu.pl/~rses/ (accesed May 15, 2010)

[30] Stanczyk, U.: On Construction of Optimised Rough Set-based Classifier. Int. Journal of Mathematical Models and Methods in Applied Sciences 2, 533–542 (2008)

[31] WHO. IPCS: Environmental health criteria 210 - Principles for the assessment of risks to human health from exposure to chemicals, Geneva (1999)

[32] Zákon č. 86/2002 Sb., o ochraně ovzduší a o změně některých dalších zákonů (zákon o ochraně ovzduší), v platném znění